DOCUMENT RESUME

ED 128 400 TM 005 564

AUTHOR Forbes, Dean W.
TITLE The Use of Rasch Logistic Scaling Procedures in the Development of Short Multi-Level Arithmetic Achievement Tests for Public School Measurement.
PUB DATE [Apr 76]
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Education; *Elementary School Mathematics; Grade 7; Grouping (Instructional Purposes); *Grouping Procedures; *Individual Differences; Mathematics; Public Schools; Standard Error of Measurement; *Student Ability; Test Construction; Testing Problems; *Test Reliability
IDENTIFIERS Rasch Item Calibration; *Rasch Model

ABSTRACT

Rasch calibration permitted the development of short achievement tests that were economical in testing time, and could be developed in a series of difficulty levels to suit student individual differences. Furthermore, these tests were of adequate reliability for practical educational measurement when individual students were assigned to tests of appropriate difficulty level. A variety of test placement strategies were considered and several were tried. Two formal procedures involving the use of a pretest screening tool for level assignment show promise of effectiveness but in the research described here tended to place many children in a test which was somewhat too difficult for them. The use of screening tests still is considered very promising although it is recommended that in the future criteria for test placement be modified so the students would be placed one, or perhaps two, levels lower than they were in the field test of these prototypes. It is further recommended that any students who get raw scores under 5, or over 25, immediately be retested with a more appropriate level to forestall the dramatic measurement error increases which occur when those limits are exceeded. (Author/BW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED128400

THE USE OF RASCH LOGISTIC SCALING PROCEDURES IN THE DEVELOPMENT OF SHORT MULTI-LEVEL ARITHMETIC ACHIEVEMENT TESTS FOR PUBLIC SCHOOL MEASUREMENT

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 564

Dean W. Forbes

Area IV
Portland Public Schools

Introduction

In the general practice of public school testing many practical problems must be dealt with. Two of the more pressing problems deal with the amount of time required for achievement testing and the difficulty of assigning each child a test with a difficulty level appropriate to his performance capability. This latter point becomes especially critical as more attention is paid to the debilitating and frustrating effects of presenting the child with a test that is far too difficult, or the motivational problems experienced by a child faced with a test which is far too easy. In both circumstances the resulting information is of very questionable validity and the emotional impact on the child is undesirable.

In an effort to improve the practical testing situation our school system, two years ago, started with an hypothetical model for an improved achievement test and began to explore various methodologies to find a procedure by which such a test could be constructed. We envisioned a test which would be parsimonious with respect to time, taking no more than 30 or approximately 40 minutes to administer. We also wanted a test which would occur in a variety of levels of difficulty within a given grade so that every child could take a test which was appropriate to his present level of functioning. This made it mandatory that there be a defensible metric underlying the multi-level tests which were planned.

Recent developments in the use of logistic functions scaling to define item and test performance, such as those presented by Rasch and championed by Dr. Benjamin Wright, seemed particularly promising. For this reason a project was undertaken to develop prototype multiple level seventh grade mathematics achievement tests to fit the model described above. The process by which the original item pool was calibrated, and by which the multiple level tests were actually developed, is documented in the appendix of the handout and will not be discussed at this time. Let it suffice, at the moment, to point out that seven short tests were developed. In local terminology they are called "level tests." Their content was arranged in ascending order of difficulty <u>within each test</u>. Average difficulty between levels (which were adjacent in difficulty) varied by a predetermined and constant interval. This meant that the level tests, themselves, formed an ascending series with respect to difficulty. They were developed within the field of general mathematics and were equally divided, in terms of content, between the conventional subtest areas of computation, concepts, and problem solving. They were administered across the seventh grade of two out of three sub-districts in the Portland, Oregon, school system (comprising in excess of 2,000 students) and the resulting data permits the study of a number of problems which had previously been identified (or which arose as the project proceeded).

At this time, preliminary information will be given with respect to a number of these problems: (1) can short tests of this nature be competitive in reliability with more traditional achievement tests? (2) is it practical, or even possible, to assign a specific child to one out of several available test difficulty levels? If such assignment to a specific difficulty level can be done, what are the most effective procedures for so doing? (3) if it is possible to devise a practical procedure for placing students in a test of appropriate difficulty level, what degree of measurement error will be involved? These problems will be discussed in sequence.

Reliability of Short Multi-Difficulty Level Tests

In discussing "level test" reliability attention will be paid to only a portion of the available data. There were seven tests built at various levels of difficulty but in the placement procedures used the student samples for levels 1 and 2 (the two easiest levels) were sufficiently small that extended analysis was not carried out. Level 7 was sufficiently difficult that it was felt less appropriate to the grade level for which the tests were developed (grade 7) than to the next grade level above.

The preliminary phases of analysis concentrated on levels 4, 5, and 6. Table 1 presents the first order Kuder-Richardson reliability coefficients for these three levels along with correlations of the level tests with an incumbent,

Table 1

Kuder-Richardson Reliabilities and Correlations with Incumbent Mathematics Achievement Tests for Level Tests 4, 5, and 6.

Level	KR_{20}	Correlation with Incumbent Test
4	.86	.46
5	.79	.72
6	.81	.75
	K = 30	

conventional mathematics achievement test. It becomes readily apparent that the reliabilities, in terms of correlation coefficient, are slightly lower than would be desired. Correlations with the incumbent test are substantial and are very consistent with that magnitude of correlation coefficient commonly found between achievement tests.

An interesting possible explanation suggests itself and is currently being studied. Despite the fact that the student's score comes from a very short test, in one very real sense the procedure under exploration logically would be analogous to taking a single long test and identifying those items of appropriate difficulty range to span the general accomplishment level of each individual student. These items are then administered to the student who is not forced to take items which are either too easy or too difficult items for him. Linking the calibrations of the various test levels into an extended scale including the items from <u>all</u> levels would make it possible to assign each student a score resulting from placement on the <u>total</u> extended scale underlying the entire series of short tests. This would be equivalent to giving a child a test of, say 150 items, and then scoring only the 30 items at the cutting edge of his performance capabilities; all easier items would (in one sense) be credited to the child's score without the necessity for the student actually to have dealt with them. If, in fact, this analog proves sound (and reliability estimates based on extended scale scores will provide one test of this in the near future) there is every reason to hope that reliability of these tests will be more than adequate for conventional educational usage.

Another way of looking at the reliability of these tests deals with the amount of error involved in a test score (since this defines the confidence band within which a given score must be interpreted). The error of measurement inherent in the calibrated scores of these short tests will be discussed in a later part of the paper but in passing it can be stated that error of measurement generally falls between one-third to one-half of a logistic scale unit and it is submitted that this is extremely competitive with many commercial tests.

Table 2 presents correlations between various parts of level tests 4, 5, and 6

Table 2

Part-Whole Correlations Involving
Level Tests 4, 5, and 6.

Level	N	Items 1-15 Correlated with Total Score	Items 16-30 Correlated with Total Score	Items 5-25 Correlated with Total Score
4	140	.88	.84	.92
5	200	.89	.83	.94
6	182	.90	.83	.96
		K = 15	K = 15	K = 20

with the totality of the same level tests. Column 1 shows the correlation between the first 15 items in the total test; column 2, between the second 15 items in the total test; and column 3, the middle 20 items with the total. Since these are part-whole correlations we would anticipate that they would be very high and for most usage they would be spurious. It is interesting to note, however, that the correlation between the middle section of the test and the total test is sufficiently high that it suggests that practically the same measurement capability exists within an even shorter test than was initially studied. This will be followed up in later research and, even if it is not utilized for measurement of individual student performance, it will have great implications for quick, easy, and painless program measurement where the goal involves the estimation of district perameters.

Strategies for Placing Students in Appropriate Test Level

A number of possible strategies were considered for placing the student in a test having an appropriate level of difficulty. Teacher assignment is, of course, possible. Another procedure would let the student examine all levels and select the one that looked the most practical to him. Level assignment could be made on the basis of the child's previous test score. Assignment could be made in terms of a short screening test which would permit identifying general level of performance capability.

One of the sub-districts involved in this research adopted this final procedure and developed a short screening test consisting of seven items, one representing the average difficulty level of items in each of the level tests. This screening test was subjected to scaleogram analysis to see if it composed a true unidimensional scale since this would provide the greatest precision in test assignment.

The test did not scale perfectly although it did have a reproducibility coefficient of .85 which was deemed adequate for practical usage. In use, the test was given to the students who then scored it themselves as the teacher read the correct answers. The student then selected the level of test to be taken in terms of number of items correct on the screening. Two variations of the strategy were used, each with a sample of the schools involved. In one group of schools the number of items correct indicated the level of the tests to be taken. Since it was felt that there was a real possibility that this would place students in a test where they could be expected to get anywhere from half to all the items correct, a second variation asked the students to take the test level that was one higher than their raw score on the screening test.

Table 3 presents numerical data about score ranges and Plate 1 graphs the range and inter-quartile range of test levels 4, 5, and 6. Data is included both for tests taken "on level" (where the screening test score indicated test level) and under "level plus one" circumstances (where the child took the next higher than was indicated by screening test score). In general, we find that placement by both strategies resulted in children taking tests somewhat more difficult than optimum.

Table 3

Ranges and Quartile Values for
Level Tests 4, 5, and 6.

Level	Test Placement "On Level"					Test Placement "Level plus One"				
	Low	Q1	Md	Q3	High	Low	Q1	Md	Q3	High
4	2	6.3	9.4	12.5	25	1	5.8	8.1	11.5	21
5	0	5.6	8.5	12.7	28	2	6.0	8.0	10.2	21
6	2	6.8	9.4	14.5	25	0	4.7	7.7	11.2	24

Generally speaking the perfect "level test", used with students that were effectively placed, would demonstrate a mean score at, or near, the mid-range of the test's raw score scale (in this case on or about 15). It would have a standard deviation such that the distribution of scores would go neither too high nor too low to provide effective measurement (i.e., the maximum raw score would not exceed 23 to 25 and the minimum not fall below 5 to 7--this would indicate a standard deviation in the vicinity of 3 raw score points). Hopefully the scores would distribute themselves fairly symmetrically although Rasch calibration makes no assumptions with respect to normal distribution. Table 3 shows that the total range of scores at every level, both for on-level and "level plus one" placement, exceeds the range considered optimum (particularly at the lower end). At most levels there are indications of a floor effect since the 25th percentile typically falls somewhere between raw score 5 and 7 with the median around 10 for on-level placement and 8 for off-level placement. The third quartile (75th percentile) falls between 12 and 14 for on-level placement and 10 and 12 for level plus one.

Examining the difference between quartile 1 and the median, and quartile 3 and the median a consistent suggestion of positive skewness emerges. This skewness, when

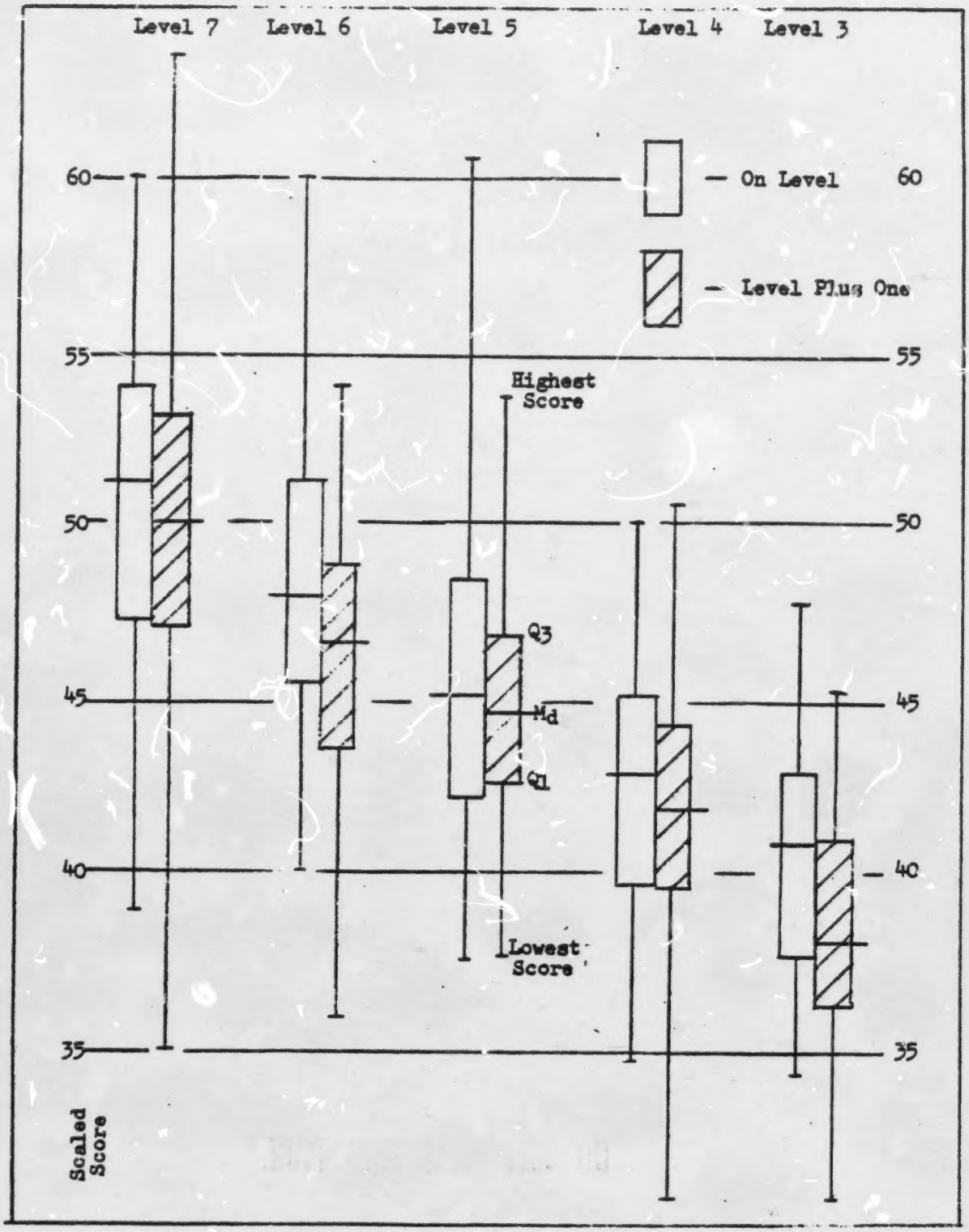

Plate 1.

Scaled Score Medians, Inter-quartile Ranges, and Total Ranges for Test Levels 3 through 7.

tested statistically, is substantial and indicates that there is a real tendency for the scores to pack up near the bottom part of the available score range with relatively few scores falling above raw score 15. (If placement were perfect, it will be recalled, the median would be approximately 15 and the quartile points would fall somewhere in the vicinity of 10 and 20.) This suggests that there is a general tendency for the two screening test procedures to place the child at a higher test level than that which is desirable for most adequate measurement.

There are many possible reasons for this and these will not be discussed at this time. (For instance, it is entirely possible that students tended to score their screening tests in such manner that they gave themselves the benefit of the doubt when the correctness of an answer was in question.) Regardless of reasons for this over-placement, the data suggests that future use of such screening tests should place the student one level lower than his raw score.

In light of the skewness which appeared in score distributions (which is not necessarily attributable to inefficiency of test level placement by means of a screening test) it becomes important to know whether or not such inaccuracies in level assignment are sufficiently great that they cast doubt upon the usefulness of the scores, and, if so, what proportion of students are so affected. It is generally acknowledged that Rasch calibration provides meaningful scores until one reaches either of the two extremes of the score scale. When a score represents nearly perfect performance, or near zero performance it is acknowledged that measurement (and logistic scaling) breaks down and no longer provides meaningful scores; this is virtually the same situation as that which occurs with any other type of test scaling since it represents the situation where you have no score due either to too high a floor or too low a ceiling (all you know is that the test did not adequately measure the student involved and you do not know how much higher or how much lower his score would have been, given an ad quate measuring instrument).

Specific criteria for upper and lower limits of effective score scale will be described in the last section of this presentation but reference to Table 4 indicates that approximately 18% of students tested on-level received raw scores below 5 whereas 24 of those tested under the "level plus one" condition fell within this range. A sufficiently small percentage under either placement strategy scored above 25 that we need not be concerned with inaccurate measurement at the upper end of the scale. The fact that so many students achieved raw scores of 5 or less raises a definite concern and will be explored in the next section of the paper.

Table 4

Percent of Cases Achieving Raw Scores below 5 on Level Tests 4, 5, and 6.

Level	Tested "On Level"	Tested "Level Plus One"
4	17	23
5	24	19
6	14	29
	$\bar{X}$ = 18.3	$\bar{X}$ = 23.7

Measurement Error

In previous sections of this paper it has been pointed out that measurement error generally is conservative for these short level tests, falling between a third and a half of a logistic scale point (which would translate to one and a half to two "standard score scale" points on the Rasch calibration scale). Table 5 presents the standard error of measurement in logistic scale terms for certain key raw score value (5, 10, 15, 20, and 25 points). Examination of the

Table 5

Standard Errors in Logistic Scale Terms
for Representative Raw Score Values of
Level Tests 4, 5, and 6.

	Raw Scores				
Level	5	10	15	20	25
4	.499	.397	.374	.396	.497
5	.497	.395	.373	.395	.496
6	.496	.395	.373	.395	.497

table shows two obvious points, that measurement error for raw scores 5 and 25 both are at or near .500 whereas the error for a raw score of 15 (the median value) is roughly .375. A second point demonstrated by the table is the fact that there is great consistency in measurement error from one level test to another among those described in this table (levels 3 through 6).

The marked similarity of data emerging in Table 5 raised the question as to how much actual difference did occur from level to level and with this in mind the curve of error relative to raw score was plotted for each level. When the error curve for each level was plotted it was found that the curves for all levels were virtually identical both in shape and in altitude. Plate 2 presents the composite curve which emerges and which is equally usable for any level from 3 through 6. An examination of this curve shows that its central portion, ro ghly from raw score 5 to 25, is relatively flat but that below 5 and above 25 it rises rapidly and symmetrically). This confirms the rule of thumb that had tentatively been adopted which was to the effect that raw scores below 5 and above 25 (for a 30 item test) should be treated with extreme caution due to probability of excessive measurement error (resulting from the fact that a test was used which was either too easy or too difficult to be valid for use with that individual). Within raw score ranges of 5 through 25 the short level tests do provide acceptable measurement, and error is within tolerable limits.

Summary

In summary it can be said that Rasch calibration has permitted the development of short achievement tests that are economical in testing time, and can be developed in a series of difficulty levels to suit student individual differences. Furthermore, these tests were of adequate reliability for practical educational measurement when individual students were assigned to tests of appropriate difficulty

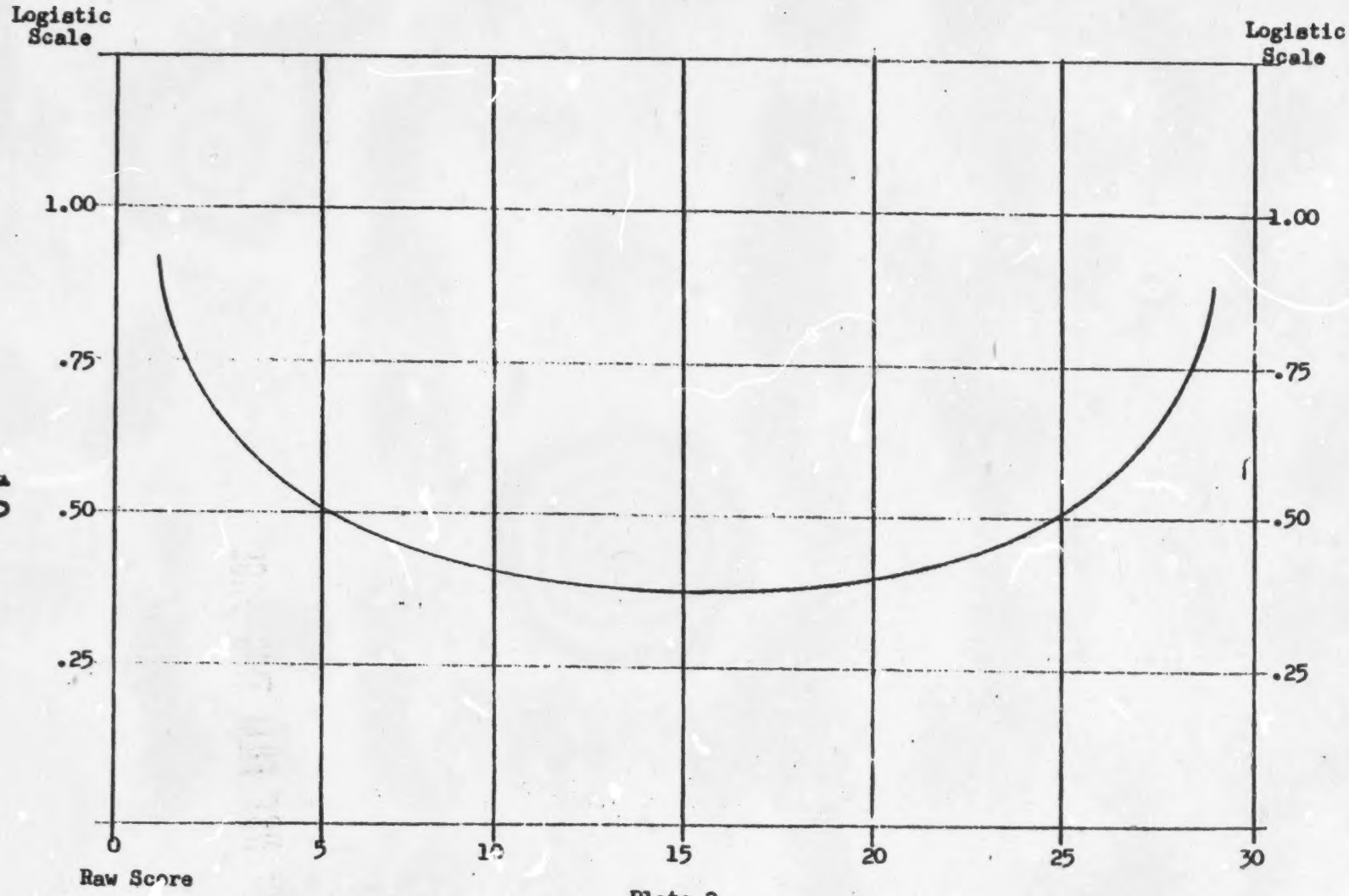

Plate 2

Curve of Standard Error as a Function of Raw Score for Level Tests 3, 4, 5, and 6.

level. A variety of test placement strategies were considered and several were tried. Two formal procedures involving the use of a pre-test screening tool for level assignment show promise of effectiveness but in the research described here tended to place many children in a test which was somewhat too difficult for them. The use of screening tests still is considered very promising although it is recommended that in the future criteria for test placement be modified so the students would be placed one, or perhaps two, levels lower than they were in the field test of these prototypes. It is further recommended that any students who get raw scores under 5, or over 25, immediately be retested with a more appropriate level to forestall the dramatic measurement error increases which occur when those limits are exceeded.

APPENDIX

Historical Documentation: Development of the Multi-Level 7th Grade Math Tests

The goal of the project was to develop a procedure for construction of better survey achievement tests that would be described by three major characteristics:

1 - shortness,

2 - relevance to instructional program, and

3 - difficulty level appropriate to present performance of each individual student.

The nature of the goals suggested the appropriateness of Rasch calibration and as a result of this the project was set up with this procedure in mind.

The resources utilized in the project were many and varied. Personnel came primarily from Areas II and III of the Portland Public School District with a great deal of consultation from the Central Evaluation Department of the Portland school system and a certain amount of involvement by the Multnomah County I.E.D. and the Metropolitan Area Testing Program. The planners and organizers of the project were the persons responsible for evaluation and measurement in their respective Areas (II and III). Data processing consultation and actual calibration runs were carried out by the Central Evaluation Department utilizing the computer center at the Bonneville Power Administration. The initial project (the prototype establishment of multi-level achievement tests, and the testing appropriateness of Rasch calibration to public school curricula) involved a file of mathematics items which were available and which had been written for 7th grade usage. This file of mathematics items had been written two years previously for the purpose of building a conventional survey mathematics achievement test for grade 7 usage and had undergone conventional, traditional item analysis. This pool was accepted as appropriate for the present project due to the nature and content of the items, to its availability, and to the fact that item analysis data was available for each single item.

Since individual items had previously been placed on cards, along with all item analysis data, the first step of the project itself was to go through the items carefully and discard any weak or ineffective items as judged by conventional item analysis. At this time, all items with difficulty levels suggesting the possibility of chance response (the very difficult items with percents pass of 20 and below) were tentatively rejected. (Further consideration suggested that items which had proved too difficult for general 7th grade usage might be very appropriate for the exceptionally able 7th grader and the 8th or 9th grader. For this reason these items were considered as a separate sub-pool but were retained in the project.)

There were 200 items in the major pool. They were arranged in order of difficulty and divided into four subsets. This gave one subset of very easy items (labeled W), one item of moderately easy items (X), one group of moderately difficult items (Y), and a group of difficult items (Z). In each case 16 to 20 items were shared between adjacent subsets (comprising the more difficult items of the easier of the two levels and the least difficult items of the other level). This was to permit the eventual linking of all four levels into one extended scale after the completion of Rasch calibration.

The difficult items were also divided into subsets with half of the more difficult items going into one test (called D-1) which shared some items with the link between W and X. (Illustration 1). The other half of the difficult items (D-3) shared items with the link between Y and Z and half of the contents of each group (D-1 and D-3) were combined with items from the X-Y link and were designated as D-2. This provided the capability of calibrating items in each of the seven subsets (W, X, Y, Z, D-1, D-2, and D-3) and also permitted linking the various subsets

together so that an extended calibration scale could be developed which would encompass the total range from the easiest item in level W to the most difficult items in the D series.

Mockup booklets were prepared for each of the seven trial tests with the previously mentioned labels (W through D-3). After editorial checking of the mockup copies, booklets and administration of the "tests" was scheduled to coincide with regular "year-end" testing programs. This would permit developing a calibration scale for each trial test, linking them all together, and developing an extended scale to tie all of the items to one continuous underlying metric.

Using the basic program for Rasch calibration developed by Wright and Panchepakesan, with modifications and improvements developed by Dr. Fred Forster of the Portland Public Schools Evaluation Department, the items in the trial tests were calibrated after administration to 7th grade students (in the case of forms W, X, Y, and Z) and unusually able 7th grade students with a mixture of 8th and 9th grade students (in the case of D-1, D-2, and D-3). Following the initial calibration, the operating characteristics of the items were re-examined and all items demonstrating weakness under Rasch calibration were eliminated (criteria for elimination of items were mean square fit in excess of 2.5, and "item-total score" correlations below .25). Approximately 25 per cent of the items in the initial pool were rejected.

Following this editing of the item pool the surviving items were reorganized, links between the various trial tests were established and verified, and one extended scale was developed to cover all of the items which survived in the total pool. At this time the total pool numbered 200 items. (The weak items which

were eliminated were balanced by the items in D-1, D-2, and D-3 which proved to be usable.

When the item pool had been refined and all weak had been deleted, it was possible to start assembling multi-level tests to fit the goals of the project. A number of questions had to be answered concerning test length, range of difficulty to be spanned by any individual test, and degree of overlap between adjacent tests. Due to lack of precedents it was necessary to arrive at some arbitrary decisions.

In order to keep testing time within reasonable limits, and to prepare packages that would fit into the conventional instructional day, it was decided to plan tests that would be manageable within one classroom period. Since the items were basically in multiple choice format this indicated a length somewhere between 25 and 40 items; a length of 30 items was selected. This would provide adequate leeway for distribution and collection time and still permit a child to attempt all of the items in the test.

The range of difficulty to be covered within any particular test presented another kind of problem since there had been no prior experience (locally or nationally) on which to build. The range of the total item pool (expressed in terms of scaled scores closely resembling conventional standard scores) was approximately 25 to 30 scaled score units. It was decided to use a test width of five score units, with an overlap of two units between adjacent levels. This permitted the development of seven different test levels ranging from very easy to very difficult.

Since the initial item pool had been described in terms of the three conventional

arithmetic subtests (computation, concepts, and problem solving) the pool was first subdivided into sub-pools representing these categories. Once the items had been listed in terms of scaled score value throughout each subtest it was possible to divide the pool into strata representing each proposed level test, identifying those items at each level which would be unique to that level as well as those that would overlap with the next higher and lower levels. With the item pool mapped in this manner it was possible to identify those specific items which would meet the criteria of calibrated scale value and content to fill the blueprint for each level test. At most levels there were more than enough items to fill the blueprint so that it was possible to select specific items within each level to provide the best balance of coverage throughout the range to be covered by that particular test. With the very lowest level there was a shortage of usable problems dealing with concepts. To fill this gap supernumerary items were selected from computation or problem solving. At the most difficult end of the scale there was a less pronounced but similar shortage in computation problems having a high degree of difficulty. With these two exceptions it was possible to fill the blueprint at each level with items that had been calibrated and demonstrated effective.

Master pages (to serve in the production of offset printing plates) were then prepared for each level with the items arranged in order of difficulty and numbered serially from 1 to 30. All pages representing one specific level were numbered in the upper right-hand corner with a large numeral indicating the test level so that there would be no confusion on the part of a student selecting the level he/she was to use. The decision was made to print all seven levels in one booklet so that the teacher would not need to worry about interfacing individual pupils with separate booklets at the right level. This represented a calculated risk because it did create the possibility that students might inadvertently take

an inappropriate level of the test. This was not seen as a serious problem due to the fact that each level was sufficiently similar to the levels above and below that any three adjacent levels should provide effective measurement for a given child and it was considered entirely possible that a child could get as much as two levels above or below that which was optimum and still be effectively measured. (It is true that with a disarticulation of two levels the child might be faced with a test that was either very difficult,or very easy, but until the point is reached where a child gets virtually a zero score on the one hand or a perfect score on the other, Rasch procedures will still develop a usable and effective score.)

Another potential problem which was carefully considered dealt with the fact that all levels would be numbered from one through 30 making it possible to use the same answer sheet for all. This made it mandatory that care be exercised in proctoring; to make sure that each child identified the test level taken on the answer sheet. Test level identification offered two major alternatives, a coding block where the child could indicate the level of the test taken (as is done in indicating grade level or sex) or a box where the child could write an Arabic numeral which was not machine scannable but which then could be transcribed in machine scannable format by a clerical operation between test administration and scoring (in the field testing stage this second alternative was selected).

At this stage of the project, seven levels of a 7th grade test existed. Each level contained 30 items equally balanced between computation, problem solving, and concepts, with the levels ascending monotonically in difficulty with substantial overlap between adjacent levels. These levels had been printed in one multilevel booklet with the pages comprising each separate level carefully and obviously identified. A machine scannable answer sheet was selected (with the same answer

sheet usable at any level of the test).

A decision was now necessary with respect to the way in which the appropriate level for any particular child could be determined. Three general procedures came to mind and were considered. These were: (1) teacher assignment to level based on judgment of pupil's present capability in mathematics; (2) arbitrary assignment of children to the median level (level 4) unless there were obvious reasons why the child should take a higher or lower level of the test; (3) assignment of test level by means of a pre-test screening device.

In Area II of the Portland, Oregon, school system where the multi-level test was field tested across the entire 7th grade (comprising roughly 1500 students in 30 elementary schools) it was decided to use a short screening test. The screening test was developed by selecting, from among the unused items in the calibrated pool, one item from the mid-difficulty range of each subtest. It was hoped that these items would form a unidimensional scale in the Guttman sense and for this reason a scalegram analysis of the performance of the items was carried out. The items did not form a true scale although the reproducability was approximately .80. Despite the fact that a unidimensional scale was not found it was decided to use the screening test by asking the children to answer the items on the test, score their own paper as the teacher read the answers, and then use their score on the screening test as an indication of which level should be taken. Two strategies were tried out, each with half of the schools in Area II. The first strategy was to have the child take the raw score on the screening test as an indication of the level test to be taken (e.g., if a child got four of the seven items on the screening test correct, the child took level 4). The second alternative was built on the rationale that a score of 4 on the screening test indicates

that the child ought to be able to handle practically everything on level 4 and, in fact, would be more appropriately served by taking level 5, one level higher than his screening test raw score would indicate. As a result, the second alternative was to use the screening test in precisely the same manner as has been previously described but then instruct the children to add 1 to their screening test score to identify the level test which was to be taken.

The field testing was carried out, the answer sheets were sorted by level of test taken and the test level coded on the answer sheet. They were then sent to the Multnomah County Intermediate Education District data processing department where they were scored by means of a digitek optical scanner and tapes were then forwarded to Dr. Forster in the Portland Public Schools Evaluation Department who performed the Rasch calibrations utilizing the data processing equipment at the Bonneville Power Agency.